Data science in beeld

Data science in beeld

Rede

uitgesproken bij de aanvaarding van het ambt van
hoogleraar Data Science for Business Analytics
aan de Faculteit Economie en Bedrijfskunde
van de Universiteit van Amsterdam
op 15 oktober 2015

door

Marcel Worring

Dit is oratie 548, verschenen in de oratiereeks van de Universiteit van Amsterdam.

Opmaak: JAPES, Amsterdam
Foto auteur: Dirk Gillissen

Mevrouw de Rector Magnificus,
Meneer de Decaan
Hoogleraren van de Universiteit van Amsterdam en zusterfaculteiten,
Waarde collegae en studenten,
Familie, vrienden,
Allen die door uw aanwezigheid blijk geven van uw belangstelling,

Vandaag wil ik het met u hebben over het nieuwe vakgebied *data science*. Een gebied dat nog zo nieuw is dat we eigenlijk nog niet eens weten wat het nou precies inhoudt. Als u hoopt na dit betoog een antwoord te hebben op deze vraag moet ik u teleurstellen. Het vakgebied moet zich nog vormen en in deze oratie kan ik slechts mijn visie laten zien en hopen dat dit ons een stapje verder brengt in dit fascinerende veld. Ik zal dit doen door u eerst mee te nemen door een klein stukje van de recente geschiedenis die heeft bijgedragen aan dit nieuwe vakgebied en kijken hoe we dit tot grote bloei kunnen brengen. Daarna wil ik specifieker inzoomen op het onderzoek dat ik zelf heb gedaan en het onderzoek dat ik binnen de leerstoel *data science voor business analytics* bij de Amsterdam Business School wil gaan ontwikkelen. Dit alles kan natuurlijk niet los worden gezien van de rest van mijn activiteiten die plaatsvinden in het Informatica Instituut. Tenslotte zal ik ingaan op de belangrijke rol van het onderwijs in data science dat moet worden ontwikkeld.

Data science komt in beeld

De digitalisering van de maatschappij in de afgelopen dertig jaar heeft onze samenleving sterk veranderd. Ik wil nu een aantal van de belangrijkste momenten van deze dertig jaar beschouwen. Het jaar 1993 was een heel belangrijk jaar, niet omdat dit het jaar was waarin ik promoveerde, maar het was het jaar waarin de Mosaic web browser werd gelanceerd door de universiteit van Illinois Urbana-Champaign. Deze browser gaf je opeens de mogelijkheid om over het Internet te surfen en is de voorganger van browsers zoals Internet Explorer en Mozilla Firefox. De kracht van de Mosaic browser was de eenvoudige visuele manier om verschillende webpagina's op het Internet te be-

zoeken. Het was slechts vijf jaar later dat een aantal slimme studenten op Stanford een nieuw bedrijf lanceerde, gebaseerd op een idee dat ze hadden om het eenvoudiger te maken om te zoeken op het Internet. Google was geboren. Het jaar 2000 bracht een andere nieuwe technologische ontwikkeling. De Sharp J-phone was de eerst mobiele telefoon met een ingebouwde camera die je in staat stelde om de foto's via de telefoon te verzenden en ze op deze manier met je vrienden te delen. Het grootschalig delen van je foto's via het Internet is de basis van Flickr dat in 2004 werd opgericht en YouTube maakte een jaar later hetzelfde mogelijk voor video's. In 2004 kwam er ook een andere grote speler in beeld namelijk Facebook dat zich richtte op het delen van je belevenissen met je vrienden. Twitter dat twee jaar later ontstond bracht weer een hele nieuwe manier van het delen van wat je bezighield. In slechts 140 tekens, een zogenaamde tweet, vertel je als auteur wat je bezighoudt en elke geïnteresseerde wereldwijd kan het lezen. Instagram dat in 2010 werd opgericht heeft een soortgelijk doel maar waar Twitter het je toelaat om eventueel een foto toe te voegen aan je tweet is het bij Instagram andersom, de foto is de belangrijkste drager van informatie. De tekst eromheen is de aanvullende informatie. Maar naast de sociale media die nu zo'n belangrijke rol hebben gekregen in onze samenleving is er de laatste jaren nog een andere belangrijke ontwikkeling te zien. Het meest actuele nieuws is te vinden op het Internet, wetenschappelijke informatie is digitaal toegankelijk, culturele instellingen maken hun collecties digitaal beschikbaar en overheidsorganisaties zoals de regering, lokale overheden en het Centraal Bureau voor de Statistiek (CBS) maken al hun informatie beschikbaar als *open data* zodat iedereen er eenvoudig toegang toe heeft. Op deze manier kunnen we bijvoorbeeld vinden welke bevolkingsgroepen in een bepaald deel van een stad wonen, maar ook wat de gemiddelde temperatuur is in Oktober. Het gevolg van dit alles is dat er in onze samenleving steeds meer data beschikbaar is.

De overvloed aan informatie heeft een impact op een heleboel verschillende domeinen. Als we bijvoorbeeld kijken naar de culturele sector zien we dat het rijksmuseum 100.000 stukken uit de collectie digitaal beschikbaar heeft gemaakt. DeviantArt (Salah et al. 2011), een platform voor online kunst, heeft maar liefst 228 miljoen kunstuitingen. De beschikbaarheid van dit soort datasets geeft een heel nieuw perspectief voor kunsthistorici. Het forensisch domein moet zich ook steeds vaker bezighouden met tekst, beeld en video. De zaak Robert M. vereiste de nauwkeurige analyse van 80.000 beelden, 800 video's en een enorm aantal tekstberichten die Robert M. met anderen had uitgewisseld. En al deze gegevens leidde weer internationaal tot nieuw materiaal om te analyseren. Laten we tenslotte kijken naar de commerciële sector. Een zoekopdracht naar een product als Nike op Instagram geeft maar liefst 33

Miljoen foto's. Bij elke foto heeft degene die de foto op Instagram heeft geplaatst een aantal zogenaamde hashtags en tekst toegevoegd. Deze geven extra informatie over wat ze aan het doen zijn, hoe ze zich voelen, of met wie ze zijn. We weten wanneer de foto is geplaatst, en vaak ook nog waar. Voor alle domeinen geldt dat ze steeds meer data gedreven worden.

U kunt zich indenken hoeveel waarde de informatie op sociale media kan hebben voor een bedrijf als Nike. Weten hoe mensen een product beleven, wat ze er over zeggen, wat ze er van laten zien op hun foto's, en hoe ze dit met anderen delen is enorm waardevol (van Dolen 2013). Het kan de optimale marketing strategie bepalen of de keuze van nieuwe productlijnen. En dat data zoveel waarde kan hebben voor een bedrijf is zo universeel dat het wel wordt gezien als de nieuwe olie. Data is echter niet alleen de nieuwe olie omdat het zo waardevol is. Net als met olie is data in ruwe vorm niet bruikbaar. Pas als olie tot benzine is verwerkt kan je er een auto mee laten rijden. Data moet eerst nog vele stappen door voordat het informatie is en waarde voor een bedrijf kan geven.

Deze waarde komt niet vanzelf. Waar de term nieuwe olie de positieve kanten benadrukt hoor je ook heel vaak de term *big data* die meestal de problemen de meeste aandacht geeft. Big data is dan iets wat dreigend en groot op je af komt. In het Engels gebruikt men daarvoor de 3 V's (Laney 2001). Big data komt in een groot volume en heeft een hoge *velocity* oftewel een hoge snelheid. Het komt doorlopend binnenstromen en kan heel snel veranderen. Tenslotte heeft big data een enorme variëteit; het kunnen getallen zijn, maar ook tijdstippen, locaties waar iets gebeurt, en teksten, beelden en video's. Al de verschillende typen van data hebben technieken nodig om te worden omgezet in bruikbare informatie.

De vraag is dus hoe we van die grote hoeveelheden data die beschikbaar zijn komen tot waarde voor een onderzoeker, organisatie of een bedrijf. En dat is in mijn ogen de kern van data science. We moeten de oplossingen vinden die ons in staat stellen om big data niet meer als een dreiging te zien maar als een waardevolle bron. We willen bereiken dat we als het ware van een afstandje rustig naar de big data kunnen kijken in al haar variëteit en dat we op die manier bewegingspatronen kunnen observeren, of de interactie tussen de verschillende informatiebronnen, waar versterken ze elkaar bijvoorbeeld en waar spreken ze elkaar tegen? Pas als we zulke patronen kunnen zien komt de volle waarde van data tot zijn recht.

Data science vereist nieuwe vormen van onderzoek

De uitdagingen waar de data science voor staat zijn groot en vereisen expertise en samenwerking die buiten de traditionele disciplines vallen en kruisverbanden nodig maken. Het is dan ook geen verassing dat we in de afgelopen twee jaar landelijk de start hebben gezien van verschillende data science centra van Eindhoven tot Groningen en van Delft tot Twente. Hoewel data in alle gevallen de basis vormt zien we toch hele duidelijke verschillen in focus. Data Science Eindhoven is bijvoorbeeld sterk in de analyse van processen zoals die te vinden is in de informatie systemen van bedrijven of uit het zogenaamde *Internet of Things,* data uit apparaatjes die met het Internet zijn verbonden. Leiden richt zich op de grondslagen van de statistiek. Dit zijn beiden initiatieven vanuit de universiteiten. Het eScience center is een NWO instituut dat landelijke ondersteuning biedt voor data gedreven wetenschappelijk onderzoek. Het Big Data Value Centrum in Almere richt zich met onder andere TNO en SURFsara op toepassingen en nauwe samenwerking met het Midden en Klein bedrijf. En de recent opgerichte Big Data Alliance is een platform dat er op is gericht om landelijke samenwerking te faciliteren tussen bedrijven en *academia*. Of het nu op basis van de wetenschappelijke inhoud of de doelgroep is, al deze partijen dragen op hun eigen manier bij aan het ontstaan van een data science landschap in Nederland. Misschien lopen ze elkaar af en toe een klein beetje in de weg, het landschap is zo breed en divers dat dit zich vanzelf zal uitsorteren. Veel interessanter is het om te kijken waar ze elkaar kunnen versterken en aanvullen.

In dit betoog zal ik mij richten op het landschap in Amsterdam en dan met name op Amsterdam Data Science waar ik in 2013 een van de initiatiefnemers van was en waar ik nu de eer heb om een *associate director* van te zijn. Amsterdam Data Science is net als de meeste andere academische data science centra een virtuele organisatie zonder een eigen gebouw, maar brengt wel honderden onderzoekers in data science bij elkaar vanuit het Centrum voor Wiskunde en Informatica, de Hogeschool van Amsterdam, de Universiteit van Amsterdam, en de Vrije Universiteit Amsterdam. Het biedt het scala van expertises die nodig zijn om de data science uitdagingen aan te kunnen gaan.

Zoals gezegd het veld is zich nog aan het vormen, ook in Amsterdam, en dus zijn er verschillende manieren om naar data science te kijken. Hier wil ik graag naar twee specifieke manieren kijken.

De wereld volgens Amsterdam Data Science is gecentreerd rond de data en heeft drie specifieke thema's. De opdeling die hier is gekozen komt voort uit de gedachte dat het belangrijk is om na te denken over hoe wetenschappers in hun eigen onderzoeksveld bepalen hoe succesvol iets is. Laten we dit nu eens

systematisch voor de vier onderdelen gaan beschouwen. Als we puur naar de data kijken is het van belang om na te denken over vragen als privacy, veiligheid en *governance* van de data. Kan ik er wel voor zorgen dat niemand achter de identiteit van een gebruiker kan komen als ik onderzoek doe op Facebook? Weet ik wel precies waar de data vandaan komt en wat er met de data is gebeurd? Is de data veilig voor aanvallen van cybercriminelen? Dit zijn allemaal vragen waar we vooral kijken of er geen fouten zitten in onze methodes en of kwaliteit gewaarborgd is. Het opslaan en verwerken van data is met name gericht op de volume en snelheid van big data. Hoe kunnen we de grenzen verleggen zodat we steeds grotere en sneller veranderende datasets kunnen gebruiken? Het analyseren en modelleren thema probeert zo goed mogelijk alle soorten van data zoals getallen, tekst en beeld te analyseren, en vaak is het doel om te kijken hoe we zo dicht mogelijk de menselijke capaciteit om dit te doen kunnen bereiken. Het modelleringsdeel van het thema probeert voorspellingen te doen op basis van de data en het doel is de voorspellingen zo goed mogelijk te laten aansluiten bij wat we daadwerkelijk in de wereld zien gebeuren. Representatie van kennis en hoe we hiermee kunnen redeneren is ook een onderdeel van dit thema omdat een kwaliteitsmaat is hoe goed het resultaat aansluit bij de kennis die we als mens hebben van de wereld. We zijn aangeland bij het laatste thema begrijpen en beslissen. We kunnen niet de data begrijpen of een beslissing nemen op basis van data als we het niet relateren aan een toepassingsdomein. Het doel is hier dus altijd gerelateerd aan iets in de buitenwereld. Wat zijn de begrippen die relevant zijn in het domein? In de forensische wereld is hier bijvoorbeeld de vraag wat is de waarde van het bewijs? In het gezondheidsdomein moeten beslissingen worden genomen zoals welk medicijn te kiezen zodat de kans op genezing het grootst is. In de kunsthistorie wordt gezocht naar antwoorden op de vraag welke schilders elkaar hebben beïnvloed. Het is dan ook hier dat we de *business analytics* plaatsen omdat de analyse van bedrijfsdata alleen maar een betekenis kan krijgen als het wordt beschouwd in de context van de doelen die het bedrijf heeft gesteld, of dat nu het maximaliseren is van de winst, verbetering van de kwaliteit, of een optimale klantbeleving. Alle thema's zijn van groot belang en hebben allemaal met elkaar te maken. Wat de thema's samenbrengt is de centrale rol die data speelt.

Laten we gaan kijken naar een andere blik op het data science landschap. In September zijn we als Amsterdam Business School begonnen met een nieuwe Master of Business Administration, oftewel een MBA, genaamd *Big Data & Business Analytics*. Daar worden ook drie thema's benoemd, maar dan vanuit de discipline waar de onderliggende technieken vandaan komen. Natuurlijk is hier een grote rol voor de bedrijfskunde weggelegd, een essentieel onderdeel

van het domein waarin de technieken die we onderwijzen worden toegepast. Als we kijken hoe we ons daar presenteren in de voorlichting is de data minder expliciet zichtbaar, maar de rol van de benodigde samenwerking tussen de verschillende disciplines komt heel duidelijk naar voren. De verschillende disciplines raken aan elkaar en hebben ook een, zij het nog kleine, overlap. Dit is een tweede kernelement van data science. Samenwerking is essentieel en speelt hier op twee manieren die ik nu verder zal uitwerken.

Om impact te kunnen hebben in een domein zal je moeten begrijpen wat er speelt in het domein en de hoofdzaken van de bijzaken kunnen scheiden. Samenwerking met experts uit het domein of het nu forensische experts, kunsthistorici of logistiek managers zijn, is de basis van elk toegepast data science project. In het zich ontwikkelende data science veld wordt dan ook vaak gesproken over T-vorm onderzoek. Kennis vanuit één specifieke discipline waarin de expertise zich jarenlang heeft opgebouwd zoals het machine leren, de econometrie, of computer visie komt samen met een applicatiegebied waarin de methode wordt toegepast. Op basis van de ervaringen worden er kleine veranderingen aan de methode gedaan om deze nog beter te maken. Als er met meerdere disciplines wordt samengewerkt aan een gemeenschappelijke toepassing spreekt men ook wel van π-vorm onderzoek. Dat wil zeggen toepassingen met diepgang in twee methodische disciplines. Graag wil ik dit illustreren aan de hand van de gele mot.

De rijk uitgedoste vleugels van de mot zijn de vele toepassingen waarin data science kan helpen. Denk hierbij bijvoorbeeld aan een marketing campagne van een bedrijf. De twee staarten symboliseren de methoden die we gebruiken, het genoemde machine leren uit de informatica, de technieken uit de econometrie, of de theorie van marketing. De grondslagen van deze methoden liggen ver uit elkaar maar in de toepassing komen ze bij elkaar en kunnen ze elkaar versterken.

Zijn er ook nog andere vormen van samenwerking? Die zijn er zeker, hoewel ze in de praktijk veel minder voorkomen. Wat nu als we ook op het methodische vlak intens met elkaar samenwerken en niet alleen in de toepassing maar ook vanaf de grondslagen van elkaar proberen te leren en elkaar proberen te versterken?

Terug naar onze mot, bij deze komen de twee staarten elkaar niet alleen maar tegen bij de vleugels, ze raken elkaar op elk punt en vormen een geheel. En dat is een sterke vorm van samenwerking. Voor de oratie was er een seminar waarin we top onderzoekers uit de informatica, econometrie, en bedrijfskunde bij elkaar hebben gebracht, hopelijk ligt hier een stap naar ook deze vorm van samenwerking. Later zal ik laten zien dat binnen mijn leerstoel, rond het onderwerp marketing, we hier al mooie stappen aan het nemen zijn.

We hebben twee kenmerken van data science besproken, de data en de samenwerking. Om tot mijn laatste punt te komen wil ik u graag meenemen naar twee beroemde onderzoekers uit het begin van mijn wetenschappelijke carrière.

De eerste onderzoeker is Herbert Freeman. Bij een van de conferenties waar ik tijdens mijn jaren als promovendus heen ging hield Freeman een *keynote*. Ik vond dat heel speciaal, want naar hem waren de Freeman codes vernoemd, een techniek in de beeldverwerking waar ik in die jaren veel mee werkte. De keynote zelf was voor mij een teleurstelling want nadat hij gesproken had over de codes die zijn naam hadden gekregen heeft hij veertig minuten lang gesproken over de gegeneraliseerde Freeman codes. Allemaal variaties op hetzelfde principe en de meerwaarde leek elke keer kleiner te zijn.

De tweede onderzoeker over wie ik wil spreken was voor mij veel meer een bron van inspiratie. Ik ben waarschijnlijk een van de weinigen in de zaal die kan zeggen dat hij Douglas Engelbart heeft ontmoet. Hij is de persoon die zowel de computer muis heeft bedacht als het idee van verschillende *windows* op je scherm. Geen van deze hebben zijn naam maar ze zijn niet meer weg te denken uit onze huidige maatschappij. Als een terzijde moet ik hier wel eerlijk zijn dat ontmoeten hier mooier klinkt dan het is. Ons gesprek vond spontaan plaats in de WC van het conferentiehotel. Zijn keynote ging verassend genoeg helemaal niet over hoe hij op het idee was gekomen van de muis, of hoe de verschillende windows op een scherm zijn ontstaan. Hij hield een overtuigend betoog dat wetenschappelijke gemeenschappen een geheugen moeten opbouwen en moeten kijken hoe ze zich als eenheid moeten ontwikkelen om tot grotere hoogte te komen. En dat is misschien wel de belangrijkste taak binnen de leerstoel die ik heb. Met slechts één dag per week kan ik als individu een klein stukje gezamenlijk onderzoek opzetten, de brugfunctie die ik heb tussen de verschillende data science disciplines is de echte waarde.

Kijkend naar het veld en naar m'n collega's op beide werkplekken zie ik enorm slimme mensen met fantastische wetenschappelijke kwaliteiten. Ik kan daarentegen ook niet ontkennen dat ik bij sommigen wel eens een sterk Freeman gevoel krijg. En we kunnen niet allemaal een Engelbart zijn, en dat moeten we ook helemaal niet willen, maar met het enorme potentieel dat er is zou een andere blik op het probleem wel eens tot nog mooiere dingen kunnen leiden. Als we niet opletten is er het grote gevaar dat data science niet meer wordt dan oude wijn in nieuwe zakken en daar is het onderwerp veel te mooi voor. Een nieuwe manier van kijken naar ons onderzoek is nodig om te komen tot een solide basis voor het zich ontwikkelende veld van data science.

Laat me het voorgaande samenvatten. Data vormt de kern van data science en het delen van datasets en hieraan werken vanuit verschillende disciplines is een uitstekende manier om samenwerking te stimuleren. Idealiter wordt deze samenwerking bereikt door van fundamenteel onderzoek tot toepassing met elkaar op te trekken. Pas als we op deze manier tot de essentie van data komen zal het nieuwe wetenschapsgebied data science tot haar volle bloei kunnen komen.

De data science van beeld

Na deze meer algemene beschouwing van het onderzoeksveld wil ik me richten op mijn eigen onderzoek zoals dat zich in de loop der jaren heeft ontwikkeld en dat nu binnen mijn leerstoel wordt voortgezet in samenwerking met de bedrijfskunde binnen de *business analytics* maar tegelijkertijd middels mijn aanstelling bij de informatica in andere domeinen zoals de kunsthistorie, de *forensics*, en de medische wereld.

Laten we teruggaan naar het jaar 2000 wat in het vakgebied dat zich bezighoudt met de analyse van grote beeldverzamelingen bekend is geworden als "het einde van de beginjaren". Het is een gevolg van de titel van het toonaangevende artikel van Arnold Smeulders waaraan ik als mede auteur een bijdrage heb geleverd (Smeulders et al. 2000). Een belangrijk begrip in dat artikel is het zogenaamde semantische gat. Waar mensen direct kunnen zien wat er op een foto staat is dat voor de computer een moeilijke taak. U zult geen moeite hebben een zebra of zonsondergang te herkennen, de computer moet het doen met slechts nullen en enen. Het dichten van het semantische gat is nu, vijftien jaar later, nog steeds een van de belangrijkste onderzoeksonderwerpen in het veld en Arnold Smeulders en Cees Snoek behoren daarin nu tot de wereldtop. Maar hoe werkt dit dan?

Het basisprincipe is eenvoudig. Je laat de computer een heleboel voorbeelden zien en dan leert de computer wat wel en niet belangrijk is. Door in dit geval foto's van een hoop zebra's te laten zien, gefotografeerd vanuit verschillende aanzichten, van dichtbij en van ver, in fel zonlicht en in de schemering leert de computer te onderscheiden wat nu wel en wat niet belangrijk is om een zebra te herkennen. Nou is dat voor zebra's en zonsondergangen vrij eenvoudig want die hebben heel herkenbare kleuren of patronen. Aan het einde van de beginjaren waren zebra's en zonsondergangen dan ook zo ongeveer de enige twee dingen die de computer wél goed kon herkennen.

U kunt zich voorstellen dat het voor een computer veel moeilijker is om, vanuit de foto's die u hier ziet, te leren hoe je een Zebu moet herkennen.

Zeker als we moeten leren hoe we de Zebu moeten onderscheiden van andere soorten runderen. In de vijftien jaar na het einde van het begin is er een enorme vooruitgang geboekt en dat komt doordat een aantal ontwikkelingen op het juiste moment bij elkaar kwamen. Reeds in het begin van de jaren negentig hadden onderzoekers methoden ontwikkeld die konden leren van beelden op een manier die erg lijkt op hoe de menselijke hersenen werken, de neurale netwerken (LeCun et al. 1990). Deze methoden waren echter in de vergetelheid geraakt, omdat ze een aantal beperkingen hadden die destijds nog niet konden worden opgelost. In het begin van deze eeuw werden er een aantal inhoudelijke successen geboekt en in 2009 werd ImageNet gelanceerd (Deng et al. 2009).

Onderzoekers in Stanford hadden voor 15.000 verschillende concepten voorbeeldfoto's verzameld en tegelijkertijd was de rekenkracht van computers zo toegenomen dat de principes van de oude, op neurale netwerken gebaseerde, leertechnieken, succesvol konden worden ingezet (Krishevsky, Sutskever and Hinton 2012). Deze technieken staan nu bekend als diep leren omdat het model dat ze gebruiken uit een heleboel lagen bestaat (leCun, Bengio en Hinton 2015). Nu is het niet zo dat de computer nu opeens al deze 15.000 concepten perfect kan herkennen maar het vormt wel een uitstekende basis voor het gebruik in nieuwe toepassingen.

Laten we een concrete toepassing beschouwen waar we in de leerstoel op dit moment naar kijken. Amsterdam promoot zichzelf al enkele jaren succesvol met de IAmsterdam-letters zoals die te vinden zijn op het Museumplein en op andere plaatsen in de stad. Een bezoek aan Amsterdam is als toerist niet meer compleet als je ook niet een selfie hebt gemaakt met de letters in beeld en deze op Instagram hebt gezet. Zoeken op Instagram naar de hashtag #IAmsterdam levert je meer dan 250.000 foto's op. Uit open data kan je bijvoorbeeld alle belangrijke attracties vinden, vervoersmogelijkheden en demografie. Voor Amsterdam marketing een waardevolle bron van informatie. Analyse van de data kan antwoorden geven op vragen als hoe toeristen denken over Amsterdam, hoe ze de letters in hun foto naar voren laten komen, en waar ze heen gaan nadat ze de letters hebben gefotografeerd en hoe ze daar komen.

Hoe komen we tot antwoorden op deze vragen? De volgende quote uit 1983 die voortkomt uit de statistiek van numerieke data, dat wil zeggen getallen, is ook voor het moderne Instagram nog steeds relevant:

> *Although we often hear that data speak for themselves, their voices can be soft and sly.* (Mosteller, Fienberg and Rourke 1983)

Een nauwkeurige analyse van de beelden en ander informatie is dus nodig. De expert kan onmogelijk alle 250.000 foto's gaan bekijken. Maar u zult zeggen we leven in de tijd van de big data en de Informatica heeft ons de tools geleverd om met grote datasets om te gaan. Daarin heeft u helemaal gelijk en we kunnen relatief eenvoudig voor alle beelden de eerder genoemde 15.000 concepten uitrekenen en een enorme hoeveelheid getallen genereren. Hier is echter een andere quote van belang die nog veel ouder is, maar ook nog steeds actueel:

> *The purpose of computing is insight, not numbers.* (Hamming 1962)

Dus aan alleen maar een grote hoeveelheid getallen hebben we niets, we zullen moeten kijken hoe we van deze getallen komen tot inzicht in de data die we tot onze beschikking hebben. Maar wat is inzicht? Het is een ongrijpbaar begrip en een exacte definitie ontbreekt. North geeft ons wel een handvat om hier over na te denken door een aantal karakteristieken van inzicht te geven (North 2006). Een erg belangrijke karakteristiek is dat het niet gebaseerd is op een enkele informatiebron, in ons geval de beelden zelf, maar op alle bronnen die beschikbaar zijn. Onze Instagram foto's hebben bijvoorbeeld geassocieerde tekst in de vorm van hashtags, een locatie waar ze zijn genomen, een tijdstip, de camera die is gebruikt en het aantal keren dat ze door anderen zijn bekeken en hoe vaak ze met een *like* zijn gewaardeerd. Pas als we deze bronnen samen beschouwen beginnen we te begrijpen waar de data over gaat. Inzicht is ook iets wat zich ontwikkeld in de tijd. De eerste keer dat we een dataset bekijken zullen we er nog niet veel van begrijpen. Op basis van wat eerste observaties kunnen we hypotheses opstellen en zo zullen we steeds verder doordringen tot alle patronen en interessante informatie die er in de dataset te vinden zijn. Door op een heleboel verschillende manieren naar de data te kijken is er een steeds grotere kans dat we resultaten vinden die we helemaal niet hadden verwacht. Maar wat de resultaten ook zijn die we vinden, ze krijgen pas waarde als ze relevant zijn binnen het domein waarin we werken. De karakteristieken maken duidelijk dat de beste manier om inzicht te krijgen een combinatie is waarin de computer en de expert samen de data analyseren. En dat is precies wat Leo Cherne in 1968 zo prachtig heeft verwoord.

> *Computers are incredibly fast, accurate, and stupid. Humans are incredibly slow, inaccurate and brilliant. The marriage of the two is beyond imagination.* (Cherne 1968)

Kijkend naar ons IAmsterdam voorbeeld, het is voor de computer geen enkel probleem om door 250.000 beelden te gaan en elke keer dat als ik dezelfde foto aan de computer laat zien zal het resultaat wat de computer uitrekent ook precies hetzelfde zijn. De expert zal op basis van voortschrijdend begrip misschien wel een steeds andere interpretatie hebben, maar zal wel subtiele elementen van de foto kunnen begrijpen. Iemand kan ongelukkig kijken op een foto, maar erbij schrijven dat het in Amsterdam fantastisch is. Is dat laatste nu sarcastisch bedoeld of juist niet? De expert zal waarschijnlijk de juiste interpretatie kiezen en zo patronen kunnen vinden die voor de computer verborgen blijven. Dus hoe kunnen we er nou voor zorgen dat de expert zo goed mogelijk met de computer kan samenwerken?

Zoals gezegd is de zoekmachine Google actief sinds 1998 en in deze zeventien jaar is het de dominante manier geworden om informatie te vinden op het Internet. In de beginjaren was het puur gericht op het vinden van tekst, maar nu kunnen we ook beelden en video's eenvoudig vinden met Google. Mijn zeer gewaardeerde collega Maarten de Rijke presenteerde in zijn oratie ZOOKMA, een utopische zoekmachine die in staat was om alles te vinden wat je maar kon bedenken (de Rijke 2006). En ik moet zeggen dat hij die utopie in de afgelopen jaren een heel stuk dichterbij heeft gebracht. Zoekmachines zijn een essentieel onderdeel van data science. Maar het nadeel van de dominantie van Google en z'n varianten is dat ze allemaal gebaseerd zijn op het simpele model waarin je een zoekvraag stelt waarna je een lijstje van antwoorden krijgt, of het nu teksten, documenten of foto's zijn. In dit lijstje staan de beste resultaten bovenaan en we kijken zelden verder dan de eerste pagina. Is dit nu de beste manier om inzicht in een grote dataset te krijgen? Soms is het voldoende, vaak ook niet. De manier hoe zoekmachines hun resultaten presenteren maakt geen gebruik van de unieke capaciteiten die de mensen heeft voor het analyseren van data.

We gebruiken dertig procent van onze hersenen om te zien (Essen, Anderson, Felleman 1992). En dit is precies wat ons unieke mogelijkheden voor het analyseren van data geeft. Het is geen toeval dat we spreken over inzicht, de patronen komen in zicht als we er goed naar kunnen kijken. Of het nu de interpretatie van een enkele foto is, of een trend in een grafiek wij komen tot resultaten die vaak beter, maar vooral aanvullend zijn aan wat een computer kan bereiken. Maar dit vereist wel dat de computer de resultaten van de analyse aan de expert toont op een manier die veel verder gaat dan een simpel lijstje. In de loop der jaren heb ik met mijn groep diverse visualisaties ontwikkeld die een expert kunnen ondersteunen in het doen van haar taak. De kracht van deze visualisaties is tweeledig. Ten eerste tonen ze de resultaten op een manier die veel meer inzicht geeft dan een lijstje. Ten tweede zijn ze uit-

gebreid met een machine leren component die het systeem in staat stelt om steeds beter te begrijpen welke beelden de gebruiker belangrijk vindt en op basis daarvan ook kan leren om de data steeds beter te analyseren. Door middel van de visualisatie kunnen de mens en de computer van elkaar leren.

Twee van deze systemen wil ik er uitlichten. De eerste is de New Yorker Melange gebaseerd op sociale media uit Foursquare, Flickr en Picasa (Zahalka, Rudinac, en Worring 2014). Dit is een systeem dat gebruikers foto's laat selecteren die ze aanspreken waarna het systeem in een aantal rondes toeristen en inwoners vindt met soortgelijke interesses. Om dit te doen maakt het systeem intelligent gebruik van state-of-the-art technieken om de foto's, teksten en locaties te analyseren. Op basis van de gevonden geestverwanten worden toeristen op attracties gewezen die ze waarschijnlijk zullen aanspreken, maar die ze zelf waarschijnlijk nooit zouden hebben gevonden. Oftewel ze hebben onverwachte resultaten gekregen in de data die heel relevant voor ze zijn op basis van verschillende bronnen. Ze hebben dat bereikt door in een aantal rondes kandidaat geestverwanten te vinden die steeds beter passen bij hun eigen profiel. We kunnen dus zeggen dat ze echt tot inzicht zijn gekomen.

De tweede zijn de multimedia *pivotables* die ontworpen zijn met in acht neming van de kenmerken van inzicht die we eerder hebben beschouwd (Worring and Koelma 2015). Ze laten combinaties zien van beeld, tekst, locatie, tijd en numerieke waarden en gebruikers kunnen daar interactief op een heleboel verschillende manieren doorheen lopen. Ze kunnen ook steeds nieuwe interpretaties van de beelden toevoegen waarvan het systeem weer kan leren. Het systeem kan gebruikers ook middels kleuren wijzen op mogelijk interessante patronen. Het is een systeem dat toepasbaar is in verschillende domeinen. Hier laat ik het IAmsterdam voorbeeld zien. Het mag duidelijk zijn dat dit een veel betere manier is om inzicht in de data te krijgen dan de ongestructureerde verzameling van beelden waar we mee zijn begonnen. Er zijn nog maar weinig systemen en methodieken die specifiek gericht zijn op het verkrijgen van inzicht in multimedia data. Met de grote toename van beeld en video in allerlei domeinen zal het belang van systemen die met multimedia kunnen omgaan enorm groeien.

Het verkrijgen van inzicht speelt zich af op de connectie tussen begrijpen, beslissen, analyseren en modeleren. Mijn onderzoek zal zich in de komende jaren met name richten op twee aspecten op deze as:

1. Hoe kunnen we de generieke methoden voor analyse en modelleren van multimedia data specifiek maken en combineren om ze te laten aansluiten bij specifieke domeinkenmerken?

2. En hoe kunnen we visualisaties ontwerpen en machine leren inzetten om deze aansluiting te doen op een manier waardoor het systeem én de mens intelligenter worden dan elk voor zich?

En daarmee kom ik terug bij het eerdere figuur en de voorbeeld applicatie. Door de data centraal te stellen en na te denken over hoe dan tot inzicht te komen is het niet mogelijk om succesvol te zijn zonder samenwerking. Om met big data om te kunnen gaan is de informatica essentieel. Omdat de expert na elke interactie meteen resultaat wil zien zal het onderzoek dat we hier voorstellen wel eens snel de grenzen kunnen bereiken van de huidige systemen. Samenwerking op het gebied van beeld- en taalanalyse, machine leren, en econometrie bieden de mogelijkheid om de grenzen van wat mogelijk is te verkennen rond de analyse en modellering. En hoe we dit alles in synergie bij elkaar kunnen brengen is ook een essentieel onderdeel van de data science waar ik in deze oratie over spreek.

Data science onderwijs

Als laatste onderwerp wil ik graag kijken naar het onderwijs. Het belang van onderwijs is evident. De nationale denktank heeft in 2014 de behoefte geanalyseerd en kwam tot de schokkende conclusie dat er in 2018 al een tekort zal zijn van 8000 hoogopgeleide data scientists (Nationale Denktank 2014). Hands-on specialisten die voor een data intensief probleem kunnen komen met een oplossing. Maar dit is slechts één groep. Er zijn daarnaast mensen nodig met heel andere profielen; de toekomstige managers die een team kunnen aansturen en de ontwikkelaars van nieuwe methodes voor analyse. We zullen dus snel moeten komen met een passend antwoord en de eerste stappen zijn genomen. De in september gestarte MBA combineert bedrijfskunde met state-of-the-art data science technieken. Deze studenten zijn pioniers en ze zullen misschien wat hinder ondervinden van het feit dat het veld zich nog moet vormen. Maar ze liggen wel op koers om in 2018 een belangrijke rol te spelen in dit nieuwe veld. Nieuwe data analyse methodieken kunnen we verwachten vanuit de econometrie en de kunstmatige intelligentie. De groep van hands-on specialisten wordt voor een deel bediend in de business analytics opleidingen aan de VU met een soort π-profiel; diepgaande kennis in de statistiek en operations research. Maar we zullen ook breed ingestelde hands-on specialisten moeten opleiden en dat is de groep waar ik nu naar wil kijken. Voor deze specifieke groep van hands-on specialisten onderwijs ontwikkelen is uitdagend omdat ze breed moeten opereren waardoor het onmogelijk is om

de diepte in te gaan op alle verschillende deelonderwerpen. Hoewel dit als een tegenstelling klinkt zal de verdieping dus moeten worden gezocht in juist dit brede aspect van de data science. Hier zal het veld zich nog verder moeten ontwikkelen.

Naar mijn mening zijn twee aspecten van belang. Als we willen kijken naar het hele proces zullen we studenten ook moeten laten werken aan het hele proces. Ik wil hierin graag de analogie maken van het produceren van een auto. Je kan studenten leren hoe ze prachtige wielen en een stuur moeten maken, maar je kan ze ook een volledige auto in elkaar laten zetten. Natuurlijk zal zo'n eerste auto niet meteen fantastisch zijn. Maar ik vind deze auto heel inspirerend. Het is een Karenjy uit Madagascar waar ze geen bloeiende auto-industrie hebben, maar wel zelf een auto hebben ontwikkeld die doet waar ie voor gemaakt is namelijk mensen van A naar B vervoeren en dat binnen de specifieke voorwaarden dat het super betaalbaar moet zijn. En de onderdelen waar over ze konden beschikken waren minimaal. Als we kijken naar data science in Amsterdam hebben we al een prachtige verzameling van onderdelen, klaar om tot iets moois samengevoegd te worden.

Het tweede aspect is de data. Uit mijn betoog mag duidelijk zijn dat data centraal zou moeten staan in een data science opleiding. En we moeten data science experts opleiden en geen experts die slechts in één domein zouden kunnen werken. Het is dus essentieel dat studenten met een rijke variëteit aan datasets in aanraking komen en dat ze zich bij de analyse daarvan volledig op kunnen laten gaan in die specifieke omgeving om precies te begrijpen wat er in dat domein speelt.

Maar net zo snel als dat ze zich hebben aangepast aan het eerste domein zullen ze weer op moeten kunnen gaan in een nieuwe omgeving. Het ontwerp van een nieuw programma dat dit mogelijk maakt zal niet eenvoudig zijn omdat het niet zomaar met bestaande vakken kan worden ingevuld. Het is in mijn ogen wel een belangrijke component van echt data science onderwijs.

Tot slot

Data science is in beeld gekomen als een nieuwe discipline en in de lijn van Leo Cherne wil ik mijn inhoudelijk deel eindigen met een eigen quote.

> *The scientific disciplines contributing to data science are scientifically sound, and narrow. Data science as a discipline is new, has limited scientific basis yet and is broad. The marriage of those disciplines with humans and massive data is beyond imagination.* (Worring, 2015)

Dankwoord

Het enige wat mij nog rest is het dankwoord. Maar als je zoals ik werkt in een brugfunctie is de lijst wel lang.

Ik wil het college van bestuur bedanken voor het in mij gestelde vertrouwen. Han van Dissel, de decaan van de Faculteit der Economie en Bedrijfskunde, en Marc Salomon als decaan van de Amsterdam Business School wil ik bedanken omdat ze mij deze prachtige leerstoel hebben aangeboden. Marc onze prettige gesprekjes over zaken als de MBA of Data Science gaan altijd gepaard met een goede kop koffie of lekker eten, je zal mij niet horen klagen. Ronald bedankt dat je me hebt opgenomen in jouw groep. Jan jij als directeur van het Informatica Instituut, en Alfons en Guus als mede directieleden bedankt voor het mogelijk maken om dit te combineren met mijn positie bij de Informatica. Met plezier zit ik nu samen met jullie in het management. Ik zie met vertrouwen de toekomst tegemoet. Amsterdam Data Science is een mooi onderdeel van de toekomst en Maarten, Sanne, Henri, en Caroline het is een voorrecht dat ik daar samen met jullie het management van mag doen. Maar het managen is natuurlijk niet de hoofdzaak. Zien dat we steeds meer mensen hierin op nieuwe onderwerpen bij elkaar brengen is de kers op de pudding. Ik zou het graag doen maar ik kan ze hier niet allemaal persoonlijk bedanken.

En dan komen we bij de grote leermeester. Arnold in 1986 begon het met een korte wetenschappelijke stage en nu meer dan dertig jaar later lopen we hier samen in toga. Wat heb ik veel van je geleerd en wat blijft het me toch verbazen dat je in alles altijd enkele jaren verder denkt dan de rest. En wat hebben we onder jouw leiding als onderzoeksgroep veel bereikt. Theo je hebt veel bijgedragen en net je eigen groep gestart en hier vorige week een oratie over gehouden. Cees, ook onze reis gaat lang terug. Je was mijn student, AIO en postdoc en nu ben je een wereldtopper. Jouw gedrevenheid en doelgerichtheid is enorm, het is fijn om tenminste voor een deel nog steeds samen met jou in ISIS te zitten. Maar geen informatica zonder systemen en Dennis, jouw nuchterheid en kennis vormen de enorm solide basis voor de groep.

Mijn huidige onderzoek is voor een groot deel tot stand gekomen met AIO's, postdocs, en andere wetenschappers. Bedankt, Xirong, Ork, Gosia, Honza, Stevan, Bjorn, en Zeno. En ook wil in het bijzonder nog Jack noemen van de TU Eindhoven. Onze gesprekken over visualisatie zijn zonder uitzondering enorm ontspannend en inspirerend.

Wat is er in de tussentijd gebeurd in samenwerking met de business school? Vlad en Noud bedankt voor onze samenwerking in het onderwijs in de econometrie. Willemijn, Bob, Masoud, met z'n vieren werken we als echt multidisciplinair team. Wat zijn de meetings gezellig en wat leren we veel van

elkaar. Hier gaan mooie dingen uitkomen! Cees en Anwar, onze eerste poging tot een gezamenlijke projectaanvraag mag dan niet gelukt zijn, er komen zeker weer nieuwe kansen.

Onderzoekers kunnen hun werk niet doen als ze niet worden geholpen met de dagelijkse zaken. Dus Brecht, Annemarie, Petra, Virginie, Atie, Sophia en alle anderen, bedankt.

Tot slot wil ik mijn familie bedanken en met name mijn moeder en mijn zus. Ze hebben me van jongs af aan in alles gesteund en er voor gezorgd dat ik ongestoord kon studeren, tonen altijd interesse en staan nog steeds voor me klaar. M'n vader, ik weet dat je heel trots zou zijn geweest.

Het belangrijkste van alles is natuurlijk mijn gezin, hoewel ze soms als ik weer eens achter de computer zit of op reis moet misschien wel eens het gevoel hebben dat dit niet zo is. Petra, Rik en Kim er is te veel om jullie voor te bedanken om het allemaal op te noemen maar één ding wil ik vandaag specifiek benoemen. Vorige week werd mij tijdens de hoogleraren introductie de vraag gesteld wat is nu privé het mooiste dat er is? Daar hoefde ik niet lang over na te denken. Voor mij is dat om samen met m'n gezin op een verre reis te zijn. Dus bedankt dat jullie met mij zulke avonturen aan willen gaan. En daarmee komen we toch weer terug bij een belangrijk thema van mijn oratie. Iedereen heeft kunnen zien hoe deze reizen een bron zijn van een grote hoeveelheid prachtige beelden.

Ik heb gezegd.

Referenties

Cherne, L. (1968), *Remarks by Leo Cherne at the Discover America Meeting,* Brussels, June 27.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), *ImageNet: A Large-Scale Hierarchical Image Database.* In IEEE Computer Vision and Pattern Recognition (CVPR).

Dolen, W. van (2013), *Zing, post, huil, tweet, lach, like en verwonder*. Inaugurele rede Universiteit van Amsterdam.

Essen, D.C. van, Anderson, C.H., Felleman, D.J., *Information Processing in the Primate Visual System: An Integrated Systems Perspective*. Science 255, p. 419-423.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012), *ImageNet Classification with Deep Convolutional Neural Networks*. Neural Information Processing Systems (NIPS).

Laney, D. (2001), *3D Data Management: Controlling Data Volume, Velocity and Variety*. Gartner.

LeCun, Y. *et al*. (1990), *Handwritten digit recognition with a back-propagation network*. In *Proc. Advances in Neural Information Processing Systems* 396-404 (1990).

LeCun, Y., Bengio, Y., Hinton, G. (2015), *Deep learning*. Nature 521, p. 436-444.

Hamming, R. (1962), preface to *Numerical Methods for Scientists and Engineers*.

Mosteller, F., Fienberg, S.E., and Rourke, R.E.R. (1983), *Beginning Statistics with Data Analysis*. Addison-Wesley Publishing Company.

Rijke, M. de (2006), *Levens zoekbaar*. Inaugurele rede, Universiteit van Amsterdam.

Nationale Denktank (2014), *Big Data in Zicht, eindrapport*.

Worring, M. (2015), *Data Science in Beeld*, Inaugurele rede, Universiteit van Amsterdam.

North, C. (2006), Toward measuring visualization insight. IEEE Computer Graphics and Applications. 26 (3), p. 6-9.

Salah, A.A.S. et al. (2011), *Explorative Visualization and Analysis of a Social Network for Arts. Journal of Convergence,* pp. 87-94.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R. (2000), *Content Based Image Retrieval at the End of the Early Years*, IEEE Transactions on Pattern Analysis and Machine Intelligence 12(22), p. 1349-1380.

Worring, M., Koelma, D.C. (2015), *Insight in Image Collections by Multimedia Pivot Tables*. In ACM International Conference on Multimedia Retrieval.

Zahalka, J., Rudinac, S. Worring, M. (2014), *New Yorker Melange: Interactive Brew of Personalized Venue Recommendation*. In ACM International Conference on Multimedia.